# PATENT ABSTRACTS OF JAPAN

(11)Publication number:

07-239799

(43) Date of publication of application: 12.09.1995

(51)Int.Cl.

G06F 11/20

G06F 3/06

G06F 13/00

(21)Application number : 06-301109

(71)Applicant: INTERNATL BUSINESS MACH

CORP (IBM>

(22)Date of filing:

05.12.1994

(72)Inventor: KERN ROBERT F

MICKA WILLIAM F MIKKELSEN CLAUS W PAULSEN MICHAEL A

SHOMLER ROBERT WESLEY

(30)Priority

Priority number: 94 199448

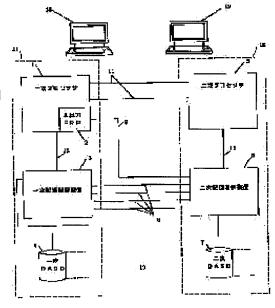
Priority date : 22.02.1994

Priority country: US

## (54) METHOD FOR PROVIDING REMOTE DATA SHADOWING AND REMOTE DATA DUPLEX **SYSTEM**

### (57)Abstract:

PURPOSE: To provide a remote data shadowing system which synchronously performs a real-time disaster recovery in a memory location base with a secondary side placed at a separated location from a primary one. CONSTITUTION: An error recovery program on a primary side 14 executes error recovery procedures and then stops the application in its execution mode to notify primary and secondary sides 14 and 15 that dual pair of faults take place. The error recovery program discriminates the cause of the dual pair of faults and restarts a dual mode if an error recovery is normally performed. When the error recovery program can not normally execute the error recovery, the update of writing on the primary side is further inhibited and an error message is sent to the operators of both primary and secondary sides 14 and 15.



(11) 特許出願公開番号 € 翐 4 盐 华 噩 么 8 (19) 日本国格許庁 (JP)

**特開平7-239799** 

(43)公開日 平成7年(1995)9月12日

噩

(51) Int.Cl.		<b>建</b> 知記中		广内整理番号	F I	技術表示
	1/20	310 A	Ą			
	3/06	304 E	臼			
1	3/00	301	Ы			

# (§ 70 諸水頂の数20 有有重新水 右

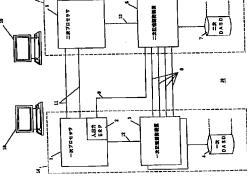
(21)出職務局	<b>特膜平6</b> -301109	(71)出職人 390009531	390009531
			インターナショナル・ビジネス・マシーン
(22) 出版日	平成6年(1994)12月5日		メ・ローボアイション
			INTERNATIONAL BUSIN
(31)優先権主張番号 199448	199448		ESS MASCHINES CORPO
(32)優先日	1994年2月22日		RATION
(33)優先權主張国	(SD) 图米		アメリカ合衆国10504、ニューヨーク州
			アーモンク(番地なし)
		(72)発明者	ロバート・フレデリック・カーン
			アメリカ合衆国85730 アリゾナ州ツーン
			ン イースト・カレヒコ・ストリート
			8838
		(74)代理人	(74)代理人 弁理士 合田 擹 (外2名)
			最然更に依く

# 強陽データ・シャドーイングを提供する方法および強陽データ二単化システム (54) [発明の名称]

(57) [要約]

同期的に記憶域ベースのリアルタイム災害復旧を行う遠 【日的】 二次側が一次側から離れた位置に配置され、 隔データ・シャドーイング・システムを提供する。

二重対の障害の原因を判別し、エラー回復が正常に行わ 一次側のエラー回復プログラムは、エラー回 **復手順を実行し、二面対の障害が発生したことを一次側** と二次側の両方に通知するために、そこで実行中のアプ リケーションを静止する。このエラー回復プログラムは れた場合は、二重モードが再開される。エラー回復プロ グラムがエラー回復を正常に実行できない場合は、一次 関での書込み更新がさらに禁止され、一次倒と二次側両 ちのオペレータに対してエラー・メッセージが送信され 【構成】



し、前記二次記憶サブシステムが入出力ERPと通信す た一次プロセッサを含み、前記一次プロセッサが一次記 ゲラム(入出力ERP)をさらに有し、二次側が二次記 **|請求項1] 一次側が一次記憶サブシステムに連結され** 憶サブシステムへの入出力書込み更新の原因となる少な 次プロセッサがそこで実行される入出力エラー回復プロ くとも1つのアプリケーションをそこで実行し、前記-記二次側が入出力哲込み更新を同期的にシャドーイング る、遠隔データ・シャドーイングを提供する方法におい 憶サブシステムに連結された二次プロセッサを含み、

(a) 街記一次記憶サブシステムから前記二次記憶サブ ンステムに入出力書込み更新を送るステップと、 (b) 前記一次および二次記憶サブシステム間に二重対 が確立されているかどうかを判別するステップと、

(c) 一枚記憶サブシステムから一枚プロセッサに障害 が発生した二重対を報告するステップと、

20 のアプリケーションから入出力ERPに一次プロセッサ (d) 二重対の障害が発生した場合に、少なくとも1つ の制御権を移転するステップと、

(e) 二重対の障害の原因を判別するステップと、

【請求項2】 二重対の状況が設定されるか回復された場 (f) 人出力ERPから二次記憶サブシステムに二重対 **衛害発生メッセージを送るステップとを含む方法。** 

【請求項3】障害が発生した「重対を報告する際に前記 ることを特徴とする、請求項1に記載の遠隔デーダ・シ 合に前記一次記憶サブシステムが更新の書込みを続行す ャドーイングを提供する方法。

8 一次記憶サブシステムが前記一次プロセッサに入出力状 兄としてチャネル終了/装置終了/装置チェックを報告 5、請求項2に記載の遠隔データ・シャドーイングを提 することがステップ (c) に含まれることを特徴とす

【請求項4】 入出力ERPが一次記憶サブシステムにセ ンス入出力を出すことがステップ(d)に含まれること を特徴とする、請求項3に記載の滋陽データ・シャドー イングを提供する方法。

9 Pによるエラー回復を実行し、二次側でデータ保全性を 【請求項5】 二重対の障害が発生した場合に入出力ER 維存するためのステップ(g)をさらに含むことを特徴 とする、諸求項4に記載の遠隔データ・シャドーイング を提供する方法。

【請求項6】エラー回復が正常に完了した後で人山力E R Pから少なくとも 1 つのアプリケーションに制御権を 返すためのステップ(h)をさらに含むことを特徴とす 5、請求項 5 に記載の遠隔データ・シャドーイングを提

に入出力ERPを介して二次側に障害発生同期遠隔コピ 【請求項7】エラー回復が正常に完了できなかった場合

**特開平7-239799** 

ー・メッセージを送るためのステップ (i) をさらに合 むことを特徴とする、請求項6に記載の遠隔データ・シ

**[糯求項8] ステップ(i)によって、一次記憶サブシ** ステムへの後続の入出力書込みがさらに防止されること を特徴とする、精求項1に記載の遠隔データ・シャドー イングを提供する方法。 [誻求項9] 一次プロセッサと二次プロセッサとの聞で 請求項 5 に記載の遠隔データ・シャドーイングを提供す エラー回復のための通信が行われることを特徴とする、

으

【請求項10】 ―次記憶サブシステムと二次記憶サブシ ステムとの間でエラー回復のための通信が行われること を特徴とする、舘求項 5 に記載の遠隔データ・シャドー イングを提供する方法。

アップのために二次側が一次側から書込み更新を受け取 に、二枚側が一枚側かの十分離れた位置に配置される遠 【鹄求項11】 一次側と二次側を有し、データ・バック 隔データニ重化システムにおいて、遠隔データニ重化シ 次側がリアルタイムのデータ可用性を提供できるよう り、一次側を使用不能にする災害が発生した場合に、

**書込み人出力更新を生成する少なくとも1つのアプリケ** 

**暦込み入出力更新を受け取って格納するための一次記憶** ーションを実行する、一次側の一次プロセッサと、

データ・リンクによって一次プロセッサに連結される、 サブシステムと、

、次側の一次プロセッサと、

一次記憶サブシステムに連結され、書込み入出力更新の 受け取ったときの特定の同期順に普込み入出力更新が書 :次プロセッサに連結され、通信リンクを介してさらに バックアップのために記憶域ベースの二重対を提供し、

結され、二重対の障害が発生した場合にエラー回復を行 **一次プロセッサ内にあって、二次記憶サブシステムに連** うためのエラー回復手段とを含む遠隔データ二重化シス き込まれる、二次側の二次記憶サブシステムと、

【請求項12】エラー回復手段が、一次プロセッサと二 次側の両方に二重対障害発生モードを報告することを特 徴とする、請求項11に記載の遠隔データ二重化システ 【請求項13】エラー回復手段が、二重対の障害の原因 ションを静止することを特徴とする、請求項11に記載 を判別しながら、一次プロセッサで実行中のアプリケー の遠隔データ二重化システム。

【請求項14】エラー回復手段が、一次記憶サプシステ ムにセンス入出力を出すことを特徴とする、請求項13

【請求項15】エラー回復手段がエラー回復を完了し、 こ記載の遠隔データニ重化システム。

二重モードが再確立された後、エラー回復手段が一次プ

20

බ

【請求項16】エラー回復手段がエラー回復を正常に完 了できない場合に、エラー回復手段が一次記憶サブシス ムの書込み入出力更新をさらに防止することを特徴と 【請求項17】 一次プロセッサに連結され、エラー回復 手段からエラー・メッセージを受け取るための一次コン する、請求項15に記載の遠隔データ二重化システム。

・メッセージを受け取るための二次コンソールとをさら こ次プロセッサに連結され、エラー回復手段からエラー に含むことを特徴とする、請求項11に記載の遠隔デー タ二重化システム。

【請求項18】 一次および二次記憶サブシステムがそれ ぞれ複数の記憶制御装置を含むことを特徴とする、請求 項11に記載の遠隔データ二重化システム。

【請求項19】通信リンクが、複数のエンタープライズ ・システム接続(ESCON)リンクを含むことを特徴 とする、請求項18に記載の遠隔データ二重化システ

ಣ

【請求項20】一次側と二次側を有し、データ・シャド ーイングのために二次側が一次側から巷込み人出力更新 に、二次側がリアルタイムのデータ可用性を提供できる ように、二次側が一次側から十分離れた位置に配置され る、記憶域ベースのリアルタイム災害復旧機能を提供す る遠隔データ「重化システムにおいて、該遠隔データー を同期的に受け取り、一次側が使用不能になった場合 重化システムが、

を提供する。

書込み入出力更新を生成する複数のアプリケーションを **書込み入出力更新を受け取るための複数の一次記憶制御 実行する、一次側の一次プロセッサと、** 

**書込み入出力更新を格納するための複数の一次記憶装置** 

データ・リンクによって一次プロセッサに連結される、 二次側の二次プロセッサと、

40 込み入出力更新をシャドーイングするために記憶域ベー リンクを介して複数の一次記憶制御装置に連結され、書 スの二重対を提供し、書込み入出力更新が生成された順 二次プロセッサに連結され、さらに複数の光ファイバ・ に提供される、二次側の複数の二次記憶制御装置と、

複数の二次記憶制御装置に連結され、そこに書込み入出 力更新を格納するための複数の二次記憶装置と、

一次プロセッサ内にあって、複数の二次記憶制御装置に **連結され、二重対の障害が発生した場合に前記複数のア** プリケーションから制御権を引き継ぎ、エラー回復を行 いながら、二重対の障害の特徴に関して一次プロセッサ および複数の二次記憶制御装置に二重対障害発生メッセ

一枚および二枚プロセッサに状 兄を出力するためのエラー回復手段とを含む遠隔データ に入出力更新を防止し、 二重化システム。

[発明の詳細な説明]

0001]

|産業上の利用分野||本発明は、一般的には、災害復旧 (DASD) データのリアルタイム遠隔コピーのための 支法に関し、より具体的には、直接アクセス記憶装置 ンステムに関する。

[0002]

用効果の高いデータ記憶を行うために、通常、複数のレ ベルに、すなわち、階層状態に分かれている。第一のレ ベル、すなわち、最高レベルのデータ記憶域は、通常は を含む。電子メモリは、ナノ秒単位でこのようなバイト 数のデータにアクセスすることによってそれぞれの回路 (従来の技術】データ処理に関連して、効率よくアクセ スし、修正および再格納できる大量のデータ(またはレ コード)を格納するためには、一般に、データ処理シス テムが必要である。データ記憶域は、効率よくしかも費 ダイナミックまたはスタティック・ランダム・アクセス ・メモリ (DRAMまたはSRAM) である電子メモリ に数百万バイトのデータを格納できる、半導体集積回路 め、このような電子メモリは最高速のデータ・アクセス の形態をとる。アクセスが完全に電子的に行われるた

0003】第二レベルのデータ記憶域は、通常、直接 ータをディスク表面に格納する。磁気DASDは、残留 アクセス記憶装置(DASD)を含む。DASD記憶域 は、磁気ディスクまたは光ディスクなどを含む可能性が あり、これらのディスクは、データのビットを構成する 「1」と「0」を表す磁気的または光学的に変質させた マイクロメートル規模のスポットとしてビット単位のデ 磁気材料で被覆した1枚または複数枚のディスクを含

8

む。これらのディスクは、保護環境内に回転式に取り付 けられている。各ディスクは、多くの同心トラックまた は間隔が詰まった円に分割されている。データは、各ト クセス機構は、一般に、1つまたは複数の読取り書込み ヘッドを含み、ディスクが回転して読取り書込みヘッド データにアクセスすることによって、DASDはギガバ イト規模のデータを格納できる。所望のデータ記憶位置 ラックに沿って1ビットずの順次格納される。ヘッド・ ディスク・アセンブリ(HDA)として知られているア を通過する際にディスクの表面にデータを転送したりデ イスクの表面からデータを転送するためにトラック間を ミリ秒単位(電子メモリより低速の単位)でこのような cディスクおよびHDAを物理的に配置する必要がある 移動できるよう、各DASDに設けられている。通常、

【0004】第二またはそれ以下のレベルのデータ記憶

20

ージを出力し、エラー回復が失敗に終わった場合にさら

ため、DASDに格納されたデータへのアクセス速度は

けではないので、次のバックアップまでにデータの一部 ては、毎晩、バックアップ場所にデータを転送する方法 コピー方法と同様、バックアップが連続して行われるわ っては受け入れられないほど、相当なデータ量が失われ プ・データを検索するのに何日もかかると同時に、数時 いう問題がある。多少改善されたパックアップ方法とし タを格納することができる。しかし、この場合も、二重 が失われる可能性がある。このため、一部のユーザにと 問分または数日分のデータが失われ、最悪の場合、同じ 災害によって保管場所も破壊されてしまう恐れがあると が考えられる。この方法では、より離れた遠隔地にデー 9 ープまたはライブラリに格納したデータへのアクセス速 **--プ記憶装置は、パックアップ日的で使用されることが 多い。 すなむむ、 路區の第二フベラバ格徴されれデータ** む。このレベルの記憶域では、必要なデータ記憶媒体を **川点は、テラバイト規模のデータ記憶など、データ記憶 塾択して装填するのにロボットが必要になるため、ライ** ブラリ内のデータへのアクセス速度はさらに低下する。 テープまたはテープとDASDライブラリを含

る恐れがある。

【0005】 データの紛失は、企業にとって破滅的なも りになる恐れがあるので、多くの企業ではバックアップ

**芰は、現在、秒単位である。** 

は、磁気テープで安全に保管するために複製される。

容量が非常に大きい割に費用が低減される点である。

のデータ・コピーを用意することが必須になっている。

一次記憶レベルで紛失したデータの復旧に要する時間

ト・プロセッサから別のホスト・プロセッサへ、または [0008] さらに最近導入されたデータ災害復旧方法 としては、遠隔方式だけでなく、連続方式でもデータが このプロセスを実現するために相当な量の制御データが 必要になる。しかし、オーバヘッドが高いために、二次 バックアップされる、遠隔二重コピーがある。あるホス 側が一次側の処理に遅れないようにする能力が妨げられ る可能性があり、このため、災害が発生した場合に二次 ある記憶制御装置から別の記憶制御装置へ、あるいはこ れらを組み合わせた方法で二重データを送信するには、 側が一次側を復旧する能力が脅かされる。

のDASDを用意する方法がある(ミラーリングと呼ば

は、追加のDASDにデータが咨き込まれるように追加 れる場合もある)。この場合、一次DASDで障害が発 生しても、データを得るために二次DASDを頼りにす

は、二重コピーが挙げられる。二重コピーの例として

も、復旧に関して考慮すべき重要な項目である。テーブ またはライブラリでのバックアップ速度の改善策として ることができる。この方法の欠点は、必要なDASDの

数が二倍になることである。

【0006】記憶装置を二重に設ける必要性を克服する もう 1 つのデータ・バックアップ方法としては、低価格 装置の冗長アレイ(RAID)構成にデータを書き込む 方法がある。この方法では、多くのDASDにデータを 割り当てるようにデータが書き込まれる。1つのDAS

リケーション・データをベースとするのではなく、汎用 ケーション・データから独立して、つまり、特定のアプ 【0009】したがって、最小限の制御データを使用し て、一次処理側のデータと一致するデータのリアルタイ ム更新を行う方法および装置を提供することが望まれて いる。この方法および装置は、復旧される特定のアプリ

(発明が解決しようとする課題】本発明の目的は、 記憶媒体をベースとして機能する。 [0010] 30

Dで障害が発生しても、残りのデータとエラー修正手順

を使用すれば、紛失したデータを復旧できる。現在で

は、数種類のRAID構成が使用できる。

【0007】一般に、記憶装置または記憶媒体で障害が 発生した場合のデータの復旧には、上記のバックアップ

側で行った哲込み更新を二次側にシャドーイング(shad owing) するための改良された設計および方法を提供す ることにある。

[0011]

SER) とDASDアドレスを持つので、これらの方法 ゲされた二次データを使用しても、システム障害の復旧

システム全体またはシステム使用現場を破壊する災害が 発生した場合にデータを復旧するには、さらに保護が必 要である。つまり、災害復旧のためには、一次データか ら離れた位置にデータの二次コピーを格納しておく必要 がある。災害保護を行う既知の方法としては、毎日また

は行えない。このため、地震、火災、爆発、台風など、

データは一次データと同じボリューム通し番号(VOL は装置障害の場合だけ有用である。しかし、ミラーリン

二次データは一次データのミラーになる、つまり、二次

方法で十分である。このようなバックアップ方法では、

更新を送信することと、一次記憶サブシステムと二次記 ッサを有する一次側を含み、一次プロセッサは、一次記 よれば、遠隔データ・シャドーイングを提供するための 憶サブシステムへの入出力害込み更新の原因となる少な くとも1つのアプリケーションを実行する。さらに一次 プロセッサは、そこで実行される入出力エラー回復プロ 側は書込み更新を同期的にシャドーイングし、二次記憶 **<b><b>∜記徴サブシステムから二次記徴サブシステムへ想込み** |課題を解決するための手段|| 本発明の第一の実施例に 方法は、一次記憶サブシステムに連結された一次プロセ **グラム(入出力ERP)を有する。二次側は、二次記憶** サブシステムに連結された二次プロセッサを含み、二次 サブシステムは入出力ERPと通信する。本方法は、 9

後サブシステムとの間に二重対が確立しているかどうか

20

プを移送する。このバックアップ計画には、バックアッ

タ位置から数キロメートル離れた安全な保管区域にテー

る。この場合、車両でテープを収集し、通常、一次デ~

は毎週、テープにデータをバックアップする方法があ

4

・バックアップのために一次側から哲込み更新を受け取 システムは、二次プロセッサに連結され、さらに通信リ **欠側に配置され、データ・リンクによって一次プロセッ** 位サブシステムは、受け取った順に害込み入出力更新が 【0012】本発明の他の実施例では、遠隔データ二重 化システムは一次側と二次側を含み、二次側は、データ る。二次側は、一次側を使用不能にする災害が発生した 場合に二次側がリアルタイム・データ可用性を提供でき るように一次側から十分離れた位置に配置される。一次 側の一次プロセッサは、書込み入出力更新の原因となる 少なくとも1つのアプリケーションを実行し、一次記憶 サブシステムは、書込み入出力更新を受け取って格納す **サに連結される。同じく二次側に位置する二次記憶サブ** ンクを介して一次記憶サブシステムに連結され、二次記 二次記憶サブシステムに書き込まれるような、書込み人 出力更新のバックアップ用の記憶城ベースの二重対を提 二重対の障害が発生した場合にエラー回復を行うために るためにそれに接続されている。二次プロセッサは、二 供する。一次プロセッサの入出力およびエラー回復は、 :次記億サブシステムに連結される。

および利点は、添付図面に図示する本発明の実施例に関 【0013】本発明の上記およびその他の目的、特徴、 する以下の詳細な説明から明らかになるだろう。 0014 [実施例] 一般的なデータ処理システムは、データを計 **算して操作し、データ機能記憶管理サブシステム/多重** 仮想記憶システム(DFSMS/MVS)ソフトウェア などを実行するために、IBMシステム/360または I BMシステム/370プロセッサなどのホスト・プロ セッサの形態をとり、少なくとも1台のIBM 399 が、メモリ制御装置と、それに組み込まれた1つまたは 5。さらに記憶制御装置は、IBM 3380または3 (DASD) に接続されている。ホスト・プロセッサが (stage) /デステージ (destage) し、変換し、全般的 0 記憶制御装置がそれに接続され、その記憶制御装置 複数のタイプのキャッシュ・メモリとを含む場合があ 390 DASDなどの1群の直接アクセス記憶装置 実質的な計算能力を提供するのに対し、記憶制御装置 は、大規模データベースを効率よく転送し、ステージ にアクセスするのに必要な諸機能を提供する。

たは遠隔地でバックアップする必要がある。一次側と二 【0015】一般的なデータ処理システムの災害復旧保 扱では、一次DASDに格独した一次データを二次側は

ップ・データ・コピーを提供するだけでなく、一次シス ない。その理由は、主に、単一の記憶制御装置では一次 **次記憶制御装置に接続された一次DASDストリングに** に接続された二次DASDストリングには二次データが **欠倒との距離は、ユーザが受け入れられる危険のレベル** こよって決まり、数キロメートルから数千キロメートル の範囲が可能である。二次側または遠隔地は、バックア テムが使用不能になった場合に一次システムの処理を引 き継ぐのに十分なシステム情報を持っていなければなら 関と二次側に設けた一次および二次DASDストリング **は一次データが格納されるのに対し、二次記憶制御装置** の両方にデータを書き込めないためである。むしろ、 格納されるのである。

アップできなくてはならない。 二次側は、一次データが 合していなければならないことである。つまり、二次デ を要する問題である。順序整合性は、それぞれが1つの データが生成され、その結果、災害復旧が崩壊する恐れ 【0016】 二次側は、一次側から十分離れている必要 があるだけでなく、一枚データをリアルタイムでバック 更新されたときに、最小限の遅延で一次データをバック アップする必要がある。しかも、二次側は、一次側で実 **行され、データまたは更新を生成するアプリケーション** ・プログラム (たとえば、IMSやDB2) を考慮せず に、一次データをバックアップしなければならない。 二 次側に要求される難しい課題は、二次データの順序が整 ータは一次データと同じ順序でコピーされなければなら ず(順序整合性)、これはシステムについてかなり考慮 データ処理システム内の複数のDASDを制御する記憶 る。順庁整合性がないと、一次データと一致しない一次 制御装置が複数存在するためにさらに複雑になってい

**つの一般的なカテゴリに分けられる。同期遠隔コピーで** 装置終了 (DE)を出力する) 前にこのようなデータの **及取りを確認する必要がある。このため、回期コピーで** は、二次側の確認を待っている間に一次DASDの入出 素である)。しかし、同期コピーは、システム・オーバ 【0017】遠隔データの二重化は、同期と非同期の2 は、一次データを二次側に送り、一次DASDの入出力 操作を終了する(一次ホストにチャネル終了(CE)と 次システムと二次システムとの距離に比例して長くなる (これは遠隔距離を数十キロメートル規模に制限する要 ヘッドを比較的小さくして、順序が整合したデータを二 力応答時間が遅くなる。一次側の入出力応答遅延は、

[0018] 非同期遠隔コピーでは、二次側でデータが (一次ホストにチャネル終了 (CE) と装置終了 (D 確認される前に一次DASDの入出力操作が完了する

E)を出力する)ため、一次側のアプリケーション・シ ステムのパフォーマンスが向上する。このため、一次D ASDの入出力応答時間は二次側までの距離に依存せ 22

9

特開平7-239799

【0021】二次ボリュームが一次ボリュームと同期し

次側を復口するために、二次個が呼び出されることがあ サブシステム、すなわち、二次記憶制御装置とDASD ていない場合、二次ボリュームは、二次システムの復旧 Rを変更する)ために非同期ボリュームを識別する必要 る。このため、二次側では、すべてのボリュームの同期 状態に関するすべての関連情報が必要である。二次記憶 は、一次側で検出された例外によって一次側が同期を解 い。たとえば、二次側が把握していない一次入出力経路 またはリンクの障害のために一次側が二次側の対等機能 と一次側アプリケーションの再開に使用することができ ない。 「次側の非同期ボリュームは非同期ボリュームと して駿別されなければならず、「次側の復旧引継ぎ手順 は、アプリケーション・アクセスを不定する(そのボリ ュームを強制的にオフラインにするか、そのVOLSE がある。一次側ホストがアクセス不能になった場合に一 除する原因となるすべての条件を判別できるわけではな にアクセスできない場合、一次側は二重対を解除するこ とがある。この場合、二次側が同期状態を示すのに対

頼りにされている場合は、始動手順によって二次DAS 状況をすべてのERP 二重対中止メッセージと組み合む センス・データがエラーの特徴を示す。このような形式 を送る。この場合、ERPの二重対中止メッセージを認 織し、その情報を二次側で確保するのは、コーザの質任 である。一次側の代わりに動作可能になるよう二次側が ソの割振りのために非同期ボリュームがオンアイン接続 システムに格納された同期状況が検索される。この同期 これは、ユーザ・システム管理機能を使用することで認 織できる。一次側の入出力操作はチャネル終了/装置終 がエラーを処理し、入出力の完了を一次側アプリケーシ ョンに通知する前に「次プロセッサに適切なメッセーツ Dが二次ホストにオンライン接続され、アプリケーショ されていないことを確認するために、二次DASDサブ アノ装置チェック(CE/DE/UC)状況で終了し、 を、外部通信によって二次側に通知することができる。 の入出力構成の場合、エラー回復プログラム(ERP) 【0022】非同期二重対ボリュームが存在すること

を含む。一次プロセッサ1は、DFSMS/MVSオペ れているアプリケーションと、システム入出力およびエ プライズ・システム/9000 (ES/9000) など 【0023】ここで図1を参照して説明すると、同図に は、一次側14と二次側15を有し、二次側15が一次 側14から20キロメートル離れている、災害復旧シス テム10が示されている。一次側14は、そこで実行さ とを有するホスト・プロセッサまたは一次プロセッサ1 レーティング・ソフトウェアを実行する I BMエンター ラー回復プログラム2 (以下、入出力ERP2という)

20

また、オンライン保守でも、同期を喪失させることがで

せると、二次側の非同期ポリューム全体を示すピクチャ

が得られる。

9 2 9 【0020】 しかし、 一次DASDが存在しアクセス可 ピーされた複数のDASDボリュームが1つのセットを 次DASDの入出力が遅延するため、非同期遠隔コピー ータの順序が一次側の更新順序と一致しなくなる場合が ステム・オーバヘッドが増加する。 したがって、一次側 で障害が発生すると、一次側と二次側との間で転送中の 災害復旧のための同期リアルタイム遠隔コピーでは、コ 形成する必要がある。さらに、このようなセットを形成 するには、各セットを構成するこれらのボリュームとそ れに対応する一次側の同等物(VOLSER)を織別す るために十分な量のシステム情報を二次側に提供する必 要がある。重要なのは、二次側が一次側と「二重対(du ができ、一次側は、二次側が同期していないことを示す ためにこのような更新にマークを付ける。二次側がいつ 内容の同期は確保されない。いくつかの理由から、二次 **側は一次側との同期性を失う場合もある。二重対が形成** プリケーションが統行できるように、二重対が中断され わけではない。この時点で保留状態の更新を適用した後 で、二次側は同期状態に戻る。一次側は、該当ボリュー って、二次側の同期を喪失させることもできる。この中 トコマンドが終了し、二重対が再確立され、保留状態の ず、一次回かの数十キロメートン籍われ遠隔站に二次回 を設けることもできる。しかし、二次側で受け取ったデ 多いので、データの順序整合性を確保するのに必要なシ ュームがこのセットと同期していない、つまり、「二重 対の障害」が発生したときに二次側がそれを認識しなけ ればならない点である。代替経路が再試行される間に一 - 次側は、二次側用の更新を待ち行列に入れる間、 - 次 則が統行できるようにコピーを中止または中断すること でも災害復旧に使用できるようにするため、二次側が一 **吹側と同期しなくなる原因になりそうな例外条件を認識** されたときに二次側は当初同期しておらず、初期データ ・コピーが完了したときに同期に違する。一次側が二次 則に更新済みデータを害き込めない場合、一次側は二重 二重対が復元されても、二次側は直ちに同期状態になる ムに関する中止コマンドを一次DASDに出すことによ より同期遠隔コピーの方が、接続障害を認識しやすい。 対を解除することもある。この場合、一次側は、更新ア plex pair)」を形成するので、1つまたは複数のボリ **能である状態で、二次側と一次側の接続を維持しても、** まり、現行の災害保護コピーを使わずに実行を続ける。 更新がコピーされると、二次側は一次側と再同期する。 することが必要である。エラー条件や復旧処置によっ た状況で一次DASDに更新内容を書き込む。このた め、二重対が復元されるまで、一次側は露出状態で、 た、一次個が一次側と一致しなくなってはならない。 【0019】 回期データ・シャドーイング データが一部紛失する恐れがある。

し、一次側は、二重対が解除されたことを示す。

またはT1/T3電話回線リンクなどの少なくとも1つ 【0025】一次プロセッサ1で実行されるアプリケー [0024] 二次側15は、チャネル13を介して1B 続されている。一次プロセッサ1は、チャネル・リンク のホスト間通信リンク11によって二次プロセッサ5に M 3990-3型などの二次記憶制御装置6に接続さ れたIBM ES/9000などの二次プロセッサ5を 含む。二次記憶制御装置6には、さらにDASD7が接 接続されている。一次プロセッサ1は、複数のエンター プライズ・システム接続(ESCON)リンク9などに よって二次記憶制御装置6との直接接続を確保すること ば、二次記憶制御装置6と通信可能になる。一次記憶制 御装置3は、複数のESCONリンクなどの複数の対等 もできる。その結果、入出力ERP2は、必要であれ 通信リンク8を介して二次記憶制御装置6と通信する。

るようになる。これに対して、書込み入出力操作が不成 れると、入出力操作が正常に完了したことを示すハード 次プロセッサ1のオペレーティング・システム・ソフト ウェアは、そのアプリケーションに普込み入出力成功の マークを付け、それにより、アプリケーション・プログ ション・プログラムによって否込み入出力操作が実行さ ラムは、最初または前の書込み入出力操作の正常終了に 依存する可能性のある次の書込み入出力操作に移行でき 功に終わった場合は、チャネル終了/装置終了/装置チ ェック (以下、CE/DE/UCという) という入山力 状況が一次プロセッサ 1 のオペレーティング・システム ・ソフトウェアに出力される。装置チェックを出力した 後、人出力ERP2は、制御権を引き継ぎ、失敗した書 **次記憶制御装置3から入手する。あるボリュームに固有** のエラーが発生した場合は、そのエラーに関連する固有 の状況が入出力ERP2に出力される。その後、入出力 ウェア状況としてチャネル終了/数隘終了(CE/D 込み入出力操作の特徴に関する具体的なセンス情報を E)が出力される。入出力操作が正常に完了すると、

特開平7-239799

一次プロセッサ1と二次 プロセッサ5との間のデータ保全性を維持するために、 新たな対等通信同期エラー回復を実行することができ との間、または最悪の場合は、

[0026] 図2および図3を参照して説明すると、同 ョン・プログラムが一次記憶制御装置3にデータ更新を 更新が一次DASD4に書き込まれ、そのデータ更新が 二次記憶制御装置 6 にシャドーイングされる。ステップ 205では、二重対の状況がチェックされ、一次側と二 **吹側が同期しているかどうかが判別される。 二重対の状** 況が同期状態になっている場合、ステップ207でデー タ更新が二次DASD7に書き込まれ、一次プロセッサ 図にはエラー回復予順が示されている。図2のステップ 201は、一次プロセッサ1で実行されるアプリケーシ 送信することを含む。ステップ203では、そのデータ 、での処理は、そこで実行されるアプリケーション・プ ログラムを介して統行される。

[0027] 二重対が「障害発生」状態になっている場 合、ステップ209で一次記憶制御装置3は、二重対で に通知する。二重対は、通信リンク8による一次記憶制 中断または障害が発生していることを一次プロセッサ1 御装置 3 と二次記憶制御装置 6 との通信障害によって

「障害発生」状態になる場合がある。あるいは、二重対

は、一次サブシステムまたは二次サブシステムいずれか 路害が通信リンク8 で発生している場合、一次記憶制御 **長置3は、二次記憶制御装置6に直接、障害を連絡する** ことができない。そこで、一次記憶制御装置3は、ステ ップ211で入出力状況としてCE/DE/UCを一次 プロセッサ1に返す。入出力ERP2は、アプリケーシ ョン・プログラムを静止させ、書込み入出力操作を要求 するアプリケーションに制御権を返す前に、エラー回復 とデータ保全性のためにステップ213で一次プロセッ のエラーによって「障害発生」状態になる場合もある。 サ1の制御権を引き継ぐ。

【0028】図3は、入出力ERP2が実行する諸ステ ップを表している。ステップ221で入出力ERP2は 力操作は、入出力エラーの原因を記述する情報を返す。 -次記憶制御装置3にセンス入出力を出す。

できる。その結果、二重対操作の現在の状況は、一次プ すなわち、このデータ記述情報は、具体的なエラーに関 して記憶制御装置または二重対の操作に固有のものにな 5。一次記憶制御装置3と二次記憶制御装置6との間の が等通信リンク8で障害が発生したことがデータ記述情 作を出す。この二次記憶制御装置6は、複数のESCO Nリンク9またはホスト間通信リンク11を介して入出 カERP2から関係ボリュームの状態を受け取ることが **報によって示された場合、ステップ223で入山力ER** P 2 は、一次記憶制御装置 3 および二次記憶制御装置 6 に対して、関係ボリュームを障害発生同期遠隔コピー状 **巻に入れるように指示する記憶制御装置レベル入出力換** 

20

ERP2は、一次記憶制御装置3と二次記憶制御装置6

特開平7-239799

406に書き込もうとするその後のすべての試みは、永 **ームに一切書き込まれなくなる。このため、必要であれ** 統エラーを受け取ることになり、対をなす二次ポリュー ムにシャドーイングできないデータは、その一次ボリュ ば、一次側アプリケーションの処置および入出力データ 操作との完全同期が可能になる。

ード(C CW))を有する一次ホスト処理エラー回復手 一夕保全性を維持する。アプリケーション・ベースのバ ステム10では、入出力命令(チャネル・コマンド・ワ イプの一次および二次サブシステム・エラーの場合にデ ックアップではなく、データ更新がリアルタイムで複写 また、災害復旧システム10は、(1)一次および二次 [0033] その結果、本明細書に記載する災害復旧シ 順によって、二重対から障害発生二重対へ一次および二 次同期遠隔コピー・ボリュームの状況を変更できるよう な外部同期遠隔コピーを取り入れ、それにより、複数タ メッセージまたはエラー・ログ記録共通データ・セット を介して具体的なポリューム更新状況に関して一次およ される昭樹梅ベースのバックアップが設けられている。 記憶制御装置ボリューム状況更新、(2) オペレータ 9

発生二重対になる場合は、一次ボリュームへのその後の (3) CRITICALボリューム識別などの、複数レ ベルの一枚/二次状況更新を試み、ボリューム対が障害 更新を防止することができる。このため、リアルタイム び二次ホスト・プロセッサが通知すること、および の完全エラー災害復旧が達成される。

【0034】 非同期データ・シャドーイング

ಜ

はそれ以上に延長できるが、複数の一次サブシステムの めに一次側と一次側との距離をさらに大きくする必要が る。一次側と二次側との距離は、現在では地球全体また 背後にある複数のDASDボリュームにわたる書込み更 新を複数の二次サブシステムに同期させることは、さら に被雑である。二次記憶サブシステム上でシャドーイン 発送することができるが、両者間でやりとりされる制御 置にわたる二次システム上でレコード書込み更新の順序 ある場合、または一次側アプリケーションのパフォーマ ンスへの影響を最小限に抑える必要がある場合に使用す ゲするために、一次データ・ムーバを介して一次記憶制 御装置から二次データ・ムーバヘレコード書込み更新を データの母は、最小限でなければならず、同時に、複数 の記憶制御装置に隠れている複数のDASDボリューム にわたる一次システムの場合と同様、複数の記憶制御装 非同期遠隔データ・シャドーインがは、1 回の災害で-次側と二次側の両方が崩壊してしまう確率を低減するた

ソフトウェアを実行するIBM ES/9000などの 【0035】図4は、一次側421と遠隔側または二次 則431とを含む非同期災害復旧システム400を示し Cいる。一次個421は、DFSMS/MVSホスト・ 一次プロセッサ401を含む。一次プロセッサ401 を正確に再構築できるものでなければならない。

欠プロセッサ 1 および二次プロセッサ 5 の向方で維持さ れる。コンソール18および19は、それぞれ一次プロ セッサ1および二次プロセッサ5からの情報をやりとり するために設けられ、入出力ERPは、両方のコンソ ロセッサ1で実行されるアプリケーションとともに、 **ル18および19に状況情報を通知する。** 

いる。このため、二次側15で復旧を試みると、二次記 「障害発生同期遠隔コピー」というマ 【0029】一次記憶制御装置3および二次記憶制御装 置6 への障害発生同期遠隔コピー入出力操作が正常に完 ムーのそのデータの状態を判別するための従来のデータ ムのデータとその同期グループ内の他のデータとの同期 **了すると、ステップ225でデータ保全性が維持されて ークを付けたボリュームを、データ回復手段(ボリュー** ベース・ログまたはジャーナル) によってそのボリュー が取られるまで使用できないものとして識別する。 僚制御装置 6 は、

RP2が受け取ったかどうかを判別するテストが行われ **--タ・セットを介して--次側15にシャドーイングされ** 【0030】ステップ227では、障害発生同期遠隔コ 制御装置6で行われた入出力操作の正常終了を入出力臣 る。正常終了すると、入出力ERP2は、ステップ22 9で一次プロセッサ1に制御権を返す。正常終了してい アーの状況 更新に つい ケー 次記 協制 御装置 3 と 二次 記憶 ない場合は、ステップ231で次のレベルの復旧通知が 行われる。この通知には、障害発生ボリュームと、一次 記憶制御装置3または二次記憶制御装置6のいずれかの を、コンソール18を介してオペレータに通知すること が含まれる。この通知は、そこで具体的なボリューム状 況を示すために、コンソール19または共用DASDデ そのボリュームの状況が正しくない可能性があること

る。エラーが修正されると、ボリューム状態は、まず保 害込み入出力操作に関する「永続エラー」回復を一次側 に回復することができる。その後、二重対が再確立され ると、データを二次DASD7に再適用することができ [0031] ステップ233で、エラー・ログ記録デー タ・セットが更新される。この更新は、一次DASD4 すると、入出力ERP2はステップ235で、障害発生 アプリケーションに実行させるために、一次側アプリケ ーションの書込み入出力操作に「永統エラー」を通知す **鉛状態(変更データの再コピー)に回復し、次に全二重** または他の記憶場所のいずれかに書き込まれ、二次側1 5 にシャドーイングされる。このエラー回復処置が完了

20 れる。CRIT=Yの場合、障害発生対の一次DASD は無関係に、一次ボリュームの永続エラー障害が報告さ 【0032】二重対を確立する場合、顧客の要求に応じ て、ボリュームをCRITICALと識別することがで きる。CRITICALボリュームの場合、ある操作の 結果、二重対の障害が発生すると、実際のエラー箇所と

でなければならない。

[0036] 一次プロセッサ401には、1BM 3990-6型記憶制御装置などの複数の一次記憶制御装置なしの複数の一次記憶制御装置かして接続されてイベ・チャネルなどの複数のチャネルをかして接続されている。また、各一次記憶測強度405には、1BM 3390Aなくとも1つのストリンがが複数されている。一次記憶制御装置405と一次DASD406によって、一次記憶サブシステムが成される。名記憶制御装置405と一次DASD406によって、一次記憶サブシステムが成される。コニットである必要はなく、両者を組み合むは、両別のコニットである必要はなく、両者を組み合むは、単別の下にてもよい。

ック407は不要である。

[0037] 一次側421から数キキロメートル離れた 体置に配置される二次面431ば、一次側421と同様 に、そこで動作する二次データ・ムーバ (SDM) 41 4を有する二次プロセッサ411を台び。二次プロセッ 4411には、当技術分野で既知の通り、光ファイバ・ チャネルなどのチャネルを介して複数の二状記憶制御数 簡415が接続されている。配徳制御装置 415には、 複数の二次DASD416を1つの制御情報DASD4 7(複数も可)が接続されている。配徳制御装置 415には、 複数の二次DASD416を1だな17によって、二次記憶サ ブンステムが構成される。

【0038】一次倒421は、過信リンク408を介して二次回431と過信する。より具体的には、一次プロセッサ401は、仮想記憶通信フクセス方式(VTAM)通信リンク408などの通信プロトコルによって、一次プロセッサ411にデータと賠額指数を情送する。

20

この通信リンク408は、縄話(T1、T3回線)、無線、無線/艦話、マイクロ波、衛星などの複数の適当な通信方式によって実現できる。

[0039]非同遊データ・シャドーイング・システム400は、一次DASD406へのすべてのデータ告込みの原序が保存され、二次DASD416に適用される(すべての一次配徳サブシステムにわたるデータ告込み(すべての一次配徳サブシステムにわたるデータ告込み個序を保持する)ように、一次配徳副御製園405から均衡データを収集する機能を含む。二次側431に渋られるデータおよび勘御構は、データ保全性を保持するのに一次側421の存在が不要になるほど、十分なもののに一次側421の存在が不要になるほど、十分なもの

[0040]アプリケーション402、403は、データまたはレコード更新を生成するが、このレコード更新 は、一次記憶制御装置405によって収集され、PDM 404によって誘み取られる。それぞれの一次記憶制御 装置405は、非同勝遠隔データ・シャドーイング・セ ッションのためにそれぞれのレコード更新をグループ化 し、非特定一次DASD406のREAD要求を介して PDM404によっちからPDM404へのレコード更新 の概述は、STARTA出力操作の回数を提供する。一次 ぶら、各一次記憶御練業留405か。分かがったッチ40 1との間で転送されるデータの金を最大にするより。 PDM404によって開御業間405と一次プロセッサ40 1との間で転送されるデータの金を最大にするように、 PDM404によって開御金数。 PDM404になっている。 PDM404によって開御金数を

[0041] データ保全性を維持しながら、PDM404がコード更新を収集し、そのレコード更新をSDM414に対してのででである。 14に対していたのでは、すべての一次配信サインステムにおいて二次DASD416に対して行われる一次DASD406のレコードWR1TEシーゲンスを再構築するのに十分な制御第一ケとともに、特定の期間の間、適切な機数の時間開係でレコードを対し、接近の数別の間、適りな機数の時間開係のレコードWR1TEシーケンスの再構築は、自己記述レコードをPDM41日から、DM41日は、所与の時間関係分のレコードが終失しているかどうか、または不完全になっているかどうかを判別するために、その日已記述レコードを検査する。

[0042] 図5および図6は、接頭部ヘッグ500(図5)と、一次記憶勘御装置405によって生成されたレコード・セット情報600(図6)とを含む、各自己記述レコードごとにPDM404が作成するジャーナル・レコード形式を示している。各自己記述レコードは、それぞれの時間関係の時間議に二次DASD416に適用できるように、それぞれの時間関係に決ちたちに

(10)

特開平7-239799

D RECORD SET機能を実行する際に (シスプ は、接頭部ヘッダ500と、各レコード・セットごとに 間報600との長さの合計を記述するための総データ長 501を含む。操作タイム・スタンプ502は、P D M タイム・スタンプである。この操作タイム・スタンプ5 レックス・クロック407に応じて)PDM404によ って生成される。一次DASD406の書込みの入出力 時間610 (図6) は、各一次記憶制御装置405のR る。操作タイム・スタンプ502は、すべての記憶制御 【0043】ここで図5を参照して説明すると、各レコ 4 0 4 が現在処理している操作セットの開始時間を示す 02は、1組の一次記憶制御装置405に対してREA SDM414に送信される実際の一次レコード・セッ EAD RECORD SETごとに固有のものであ ード・セットの先頭に挿入される接頭部ヘッダ500 ETコマンドは、PDM404によって出されるが、 装置で共通のものである。READ RECORD 下の条件のいずれかの場合に予測できる。

(1) 一次記憶制御装留405の所定のしまい値に基ろく、その一次記憶制御装置のアナンション割込み(2)所定の時間関係に基ろく、一次プロセッサ401のタイを割込み

(3) レコード・セット情報が、使用可能であるがまだ 窓み取られていない未解決のレコード・セットに関する 道加情報を示す場合

条件(2)では、タイマ問隔を使用して、低レベル活動の期間中に二次システム431がどの程度遅れて実行するかを崩御する。条件(3)は、PDM404がご一次記憶制御装置405の活動に遅れないようにするためにさらに活動を駆動する処理問隔中に、PDM404がすべてのレコード・セットを待ち行列処理しなかった場合に発生する。

ープ503内の各レコード・セットバとに一次記憶制御 レコード読取り時間507は、すべての一次記憶制御装 置405に共通の操作タイム・スタンプを提供し、現行 [0044] 時間間隔グループ番号503は、現行レコ ド・セット (整合性グループのうちの所与の時間間隔 グループにしい たすべたの一次 記憶制御装置 405 にむ スタンプ502とレコード読取り時間507によって境 る。グループ内順序番号504は、所与の時間間隔グル 装置405用のアプリケーションWR I TE入出力の順 序を(PDM404に対して)織別するためにハードウ ェアが提供するIDである。一次SSID(補助記憶域 たるレコードのセット)が風す時間間隔(操作タイム・ 界が示される)を識別するために B D M 4 0 4 が出力す ID)505は、各レコード・セットゴとに一次記憶制 御装置405の特定の一次記憶制御装置を明確に識別す パフォーマンス上の考慮事項に応じて、P DM404ま るものである。二次ターゲット・ボリューム506は、 たはSDM414のいずれかによって割り当てられる。

【0045】次に図6を参照して説明すると、レコード 有情報601~610は、レコード更新が行われた実際 号 (CCHH) 602は、各レコード更新ごとの一次D ションIDである一次SSID603は、一次SSID 505と同じものである。状況フラグ604は、特定の 祝情報を提供する。順序番号605および630は、レ 込み操作のタイプを職別する操作標識である。この操作 クタを職別する。カウント・フィールド609は、後続 される。特定のレコード・データ620は、各レコード ドを提供する。最後に、順序番号630は、読み取られ ・セット情報600は、一次記憶制御装置405によっ て生成され、P DM404によって収集される。更新固 の一次DASD406を含む各レコードの一次装置ユニ ット・アドレス601を含む。シリンダ番号/ヘッド番 ゲータ・レコード620が後に続くかどうかに関する状 のデータ)が読み取られたかどうかを示すために各レコ ードに番号を1つずつ割り当てる。一次DASD書込み 入出力タイプ606は、各レコードについて行われた書 ク・レコード・フォロー、完全トラック・データ・フォ 検索引数607は、最初に読み取られたレコード・セッ ト・データ・レコード620に関する初期位置決め情報 を示す。セクタ番号608は、レコードが更新されたセ の特定のレコード・データ・フィールド620の数を記 述する。一次DASD406の書込み更新が行われたホ スト・アプリケーション時間は、更新時間610に記録 更新ごとのカウント/キー/ゲータ(CKD) フィール たレコード・セット全体がPDM404に転送されたか ASD406上の位置を示す。一次記憶制御装置のセッ コード・セット全体(PDM404に転送されたすべて 標識は、更新書込み、フォーマット書込み、部分トラッ ロー、消去コマンド実行、または全書込み実行を含む。 どうかを示すために順序番号605と比較される。 問隔のレコード・セットの終了時間を示す。 8 20

ットにわたる) は、操作タイム・スタンプ502、時間 7、一次装置アドレス601、一次SSID603、お 新をコピーできるように、ソフトウェア・グループが呼 整合性グループを作成するのに使用する情報(すべての 次SSID603を含む。完全復旧可能な一次DASD [0046] ―次DASD406でレコード更新が書き 込まれたのと同じ順序でSDM414がそのレコード更 記憶制御装置405から収集したすべてのレコード・セ 間隔グループ番号503、グループ内順序番号504、 -|次制御装置SSID505、レコード読取り時間50 用のすべてのレコードがSDM414側で各記憶制御装 **置405ごとに受け取られたかどうかを判別するのに使** 用する情報は、時間間隔グループ番号503、グループ 内順序番号504、物理制御装置1D505、および一 406のレコード更新と同等に二次DASD416上に び出した整合性グループで更新レコードが処理される。 よび状況フラグ604を含む。1つの時間間隔グループ 20

20 たはSDM414のいずれかが収集した構成情報を提供 隔コピーは、ここに記載するのと同じ非同期遠隔コピー 【0047】図7および図8は、仮旧時間とジャーナル **耺送時間を単純化した現行ジャーナル内容を記述するた** めの状態テーブル700とマスタ・ジャーナル800を それぞれ示す。状態テーブル700は、PDM404ま 号)およびその制御装置でのボリュームと、対応する二 **外記憶制御装置のセッション I Dおよび対応するボリュ** ームとを含む。このため、構成情報は、どの一次ボリュ ーム710または一次DASDエクステントが二次ボリ ューム711または二次DASDエクステントにマッピ ングされるかを追跡する。状態テーブル700まで単純 に拡張して部分ボリューム・エクステント712 (CC HHからCCHHまで)を示す場合、部分ボリューム遠 方法を使用して達成できるが、完全ボリュームの場合よ り細分性(トラックまたはエクステント)はより細かく し、一次記憶制御装置のセッションID(SSID番

システム410がもはや存在しないスタンドアロン環境 ープ番号、ジャーナル・ボリューム上の位置、および操 作タイム・スタンプを含む。また、マスタ・ジャーナル 【0048】マスタ・ジャーナル800は、整合性グル 800は、整合性グループにグループ化した特定のレコ **ード更新を維持する。状態テーブル700とマスタ・ジ** ャーナル800は、災害復旧をサポートするため、一次 で動作できなければならない。

【0049】制御項目全体が正しく書き込まれるように ル800の前後に置かれる。このタイム・スタンプ制御 は、さらに二次DASD417に寄き込まれる。制御要 寮は二重項目 (1) および (2) を含み、次に示す例の するため、タイム・スタンプ制御は各マスタ・ジャーナ ように一方の項目が必ず現行項目になる。

- (1) タイム・スタンプ制御 | 制御情報 | タイム・スタ
- (2) タイム・スタンプ制御 | 制御情報 | タイム・スタ

いかなる時点でも、(1)または(2)のいずれかの項 しいタイム・スタンプ制御を持つ項目である。災害復旧 5にどのレコード更新が適用されたかを判別するのに使 目が現行または有効項目になるが、有効項目は前後に等 では、制御情報を得るために、最新のタイム・スタンプ を持つ有効項目を使用する。この制御情報は、状態情報 (記憶制御装置、装置、および適用される整合性グルー プに関する環境情報)とともに、二次記憶制御装置41

所定の時間間隔の間にすべての一次記憶制御装置405 こわたるすべての読取り レコード・セットが二次側43 で受け取られると、SDM414は、受け取った制御 情報を解釈し、レコード更新が最初に一次DASD40 6 上で書き込まれたのと同じ順序でそのレコード更新が **適用されるように、受け取った諮取りレコード・セット** をレコード更新・グループとして二次DASD416に 適用する。このため、一次側アプリケーションの順序

れる。このプロセスは、以下、整合性グループの形成と 込みが制御装置の順序命令に違反しない場合は、どのよ "を受け取る前に従属書込み番号2を実行することがで スタンプを持つ第一の書込みと同じレコード・セット整 (データ保全性) 整合性はすべて二次側431で維持さ **呼ぶ。整合性グループの形成は、次のような仮定に基づ** いて行われる。(A)独立しているアプリケーション書 (B) 従属しているアプリケーション書込みは、タイム ・スタンプの順に実行しなければならないため、アプリ アーションは、書込み番号 1 から制御装置終了、装置終 きない。 (C) 第二の弘込みは必ず (1) 遅いタイム・ 合性グループに入るか、 (2) 後続のレコード・セット うな順序でもアプリケーション書込みを実行できる。 整合性グループに入る。

[0051] 図9を参照して説明すると、同図には、記 整合性グループの形成例(整合性グループは一次側42 は昇順に発生するものと想定する。 時間間隔T1の操作 筬制御装置SSID1、SSID2、およびSSID3 が、この例では明解にするため3つ使用する)に関する 1または一次側431のいずれにも形成できるはずであ る) が示されている。時間問隔T1、T2、およびT3 る。PDM404は、時間関隔T1~T3の間に記憶制 御装置SS1D1、2、および3からレコード・セット プ1であるG1 (時間西隔グループ番号503) に割り タイム・スタンプ502は、記憶制御装置SSID1、 ・データを入手する。時間間隔T1のSSID1、2、 および3に関するレコード・セットは、時間間隔グルー など(記憶制御装置はいくつでも含めることができる SSID2、およびSSID3について設定されてい 当てられる。グループ内順序番号504は、SSID 1、2、および3のそれぞれについた示され、この場 合、SSID1は11:59、12:00、および1 40 3

1:58、11:59、および12:02に3つの更新 を持つ。時間関隔T2およびT3のレコード・セットは 2:01に3つの更新を待ち、SSID2は12:00 および12:02に2つの更新を待ち、SSID3は1 川挙されているが、衛略化のため、更新時間の例は示さ れていない。 【0052】ここで、二次側431で受け取った制御情 報およびレコード更新に基づいて、整合性グループNを 生成することができる。時間間隔グループ番号1のレコ

20

特開平7-239799

とに一次記憶制御装置405によってフコード・セット 20で各一次記憶制御装置405との遠隔データ・シャ は、データまたはレコードがシャドーイングされる一次 ボリュームの識別を含む。ステップ1030では、各ア ドーイング・セッションを開始するが、このステップ プリケーションWRITE入出力操作(図6を参照) 情報600がトラッピングされる。

ション・メッセージを含むプロンプト、所定のタイミン PDM404が、捕捉したレコード・セット情報600 ステップ1050でPDM404がレコード・セットの 読取りを開始すると、PDM404は、各レコード・セ レコードは、接頭部ヘッダ500と、レコード・セット **情報600を含む)を作成するための接頭部ヘッダ50** 0 (図5を参照) を付ける。このジャーナル・レコード 【0055】ステップ1040は、前述の通り、アテン ットの控行 条匠のジャーナヴ・フロード (ジャーナヴ・ には、二次側431(または一次側421)で整合性グ グ間隔、または読取りレコード数増加の通知に応じて、 を各一次記憶制御装置405から読み取ることを含む。 ループを形成するのに必要な制御情報(およびレコー

ド)が含まれる。

ている限り、任意の順序にすることができる。

[0056] ステップ1060では、PDM404が通 る。SDM414は、ステップ1070で状態テーブル 号別および順序番号別に収集する。ステップ1080で SDM414は、ジャーナル・レコードを検査し、各時 ル・レコードが完全な場合は、SDM414による整合 される場合は、同じデータ・ムーバ・システム内で)S 置405ごとに、受け取ったレコード更新をグループ番 間間隔グループごとにすべてのレコード情報を受け取っ たかどうかを判別する。ジャーナル・レコードが不完全 な場合は、ステップ1085によって、SDM414は う通知する。PDM404が正しく再送信できない場合 信リンク408を介して(整合性グループがそこで形成 700を使用し、データ・シャドーイング・セッション 用に確立した各時間間隔グループおよび一次記憶制御装 P DM4 0 4 に必要なレコード・セットを再送信するよ は、二重ボリューム対に篩害が発生している。ジャーナ 性グループの形成を含むステップ1090が実行され DM414に生成したジャーナル・レコードを送信す

**次DASD417 (図4) 上のSDM414ジャーナル** 0は、時間間隔グループが完全かどうかを判別するテス ループを形成するためのステップ1090 (図10) を 表すステップ1100~1160が示されている。整合 ログ ("hardened") に書き込まれる。ステップ111 は、少なくとも1つの読取りフロード・セット・バッフ 【0057】図11を参照すると、同図には、整合性グ このステップでは、各ソフトウェア整合性グループが二 トを実行する。すなわち、各一時記憶制御装置405 性グループの形成は、ステップ1100から始まるが、

8

る可能性がある。

(12)

2 ク407の十分な解像度が与えられる可能性はほとんど ード更新が時間間隔グループ番号2のレコード更新より が設定される。この例では、最小時間は12:01にな る。最小時間と等しいかそれ以上の諮取りレコード・セ N+1に含まれる。1つのボリュームに対する2つのレ ないが、時間間隔グループN内の早い順序番号を持つレ コード更新は、整合性グループN用のそのグループとと もに保管される。ここで、レコード更新は、読取りレコ ード・セット時間に基心にて順序のけされる。複数のフ コード更新の時間が等しい場合、小さい順序番号を持つ レコード更新は、大きい順序番号を持つレコード更新の 前に置かれる。これに対して、複数のレコード更新の夕 は、そのレコード更新が同じ整合性グループに保管され 1、2、および3のそれぞれの最後のレコード更新の最 も早い読取りレコード・セット時間と等しい、最小時間 ット時間を有するレコード更新はすべて整合性グループ コード更新時間が等しい場合、シスプレックス・クロッ 堅くならないようにするため、記憶制御装置SS1D イム・スタンプが等しいが、ボリュームが異なる場合

合性グループの形成では、一次記憶制御装置405に対 する書込み操作にタイム・スタンプが付けられると予想 【0053】 一次記憶制御装置405が、指定の時間間 隔の間に読取りレコード・セットへの応答を完了しなか で、整合性グループを形成することはできない。 一次記 **末着割込みのために、システムの末着割込みハンドラが 吹記憶制御装置405が適切な時間に操作を完了した場** 合は、入出力が完了に至り、通常操作が総行される。整 される。しかし、プログラムによっては、タイム・スタ の場合、一次記憶制御装置405は、タイム・スタンプ としてゼロを返す。整合性グループの形成は、データが **読み取られたタイム・スタンプに基づいて、タイム・ス** タンプを持たないこれらのレコードの境界を示すことが できる。 整合性グループの時間別にレコード更渐の境界 を容易に示せないほど、タイム・スタンプを持たないレ 二重ボリュームが同期していないというエラーが発生す コード更新が一定の時間間隔の間に多数発生した場合、 **憉制御装置405がその操作を完了しなかった場合は、** った場合、その一次記憶制御装置405が完了するま 制御権を受け取り、操作が終了する。これに対して、 ンプが付けられずに毎込みが生成されるものもある。

20 スタンプが付けられる。PDM404は、ステップ10 形成する方法を示す流れ図である。図10を参照して説 男すると、このプロセスは、ステップ1000から始ま ス・クロック407を同期クロック(図4)として使用 [0054] 図10および図11は、敷合柱グループを ングを確立する。ステップ1010では、シスプレック して、すべてのアプリケーション入出力操作にタイム・ り、一枚側421が、行うべき脳隔ゲータ・シャドーイ

または複数の二重ボリューム対に障害が発生している可 を提示したか、レコード・セット・バッファ内にこの ようなレコード更新が置かれていないという確認をPD (またはヌル) を拵しすべての読取りフョード・セット ・バッファがSDM414によって受け取られていなけ ればならない。時間間隔グループが不完全な場合は、ス 一次記憶制御装置405からのレコード・セットの読取 りを再試行する。エラーが発生した場合は、特定の1つ 能性がある。完全な時間間隔グループを受け取ると、ス テップ1120は、第一の整合性グループ・ジャーナル ・レコードを判別する。この第一(または現行)の整合 性グループ・ジャーナル・レコードとは、最も早い操作 タイム・スタンプ502と、同じ操作タイム・スタンプ 502を持つすべてのレコードの最も早い更新時間61 テップ1110は、必要なデータが受け取られるまで、 M404から受け取らなければならず、しかもデータ 0 とを含むレコードである。

て、どのレコードがそこに最後に含めるレコードかを判 プ・ジャーナル・レコードに含まれる)。 現行整合性グ ループ・ジャーナル・レコードの最後のレコードは、各 一次記憶制御装置405ごとに最大更新時間のうちの最 一次記憶制御装置405の最後の更新が比較され、これ ちのうちの最も早いものだけが現行整合性グループ・ジ 【0058】ステップ1130は、現行整合性グループ 別する(一部のレコードは除去され、次の整合性グルー 小更新時間(最小時間)として判別される(つまり、各 ・ジャーナル・レコードに含まれるレコードを検査し ヤーナル・レコードに残る)。

ドに残っているこれらのレコード更新は、ステップ11 40で、更新時間610とグループ内順序番号504に **応じて順序付けされる。レコード更新を特たない一次記** 憶制御装置405は、整合性グループに関与しない。ス テップ1150では、現行整合性グループの残っている プ内順序番号504は、空バッファで終わり、その操作 507および更新時間610と併せ使用して、一次記憶 制御装置405におけるアプリケーションWRITE入 時間間隔の間にかくての概要のフロード・セットが眺み 取られたことを示すはずである。空バッファがない場合 は、現行ソフトウェア整合性グループ内の最後のレコー ドを定義するステップ1120を、レコード読取り時間 【0059】現行整合性グループ・ジャーナル・レコー が、次の整合性グループに渡される。それぞれのグルー レコード更新 (最小時間より選い更新時間を持つもの) 山力操作の正しい順序を決定することができる。

[0060] ステップ1160は、完全災害復旧の制約 の下で特定の書込み更新が二次DASD416に適用さ エンドを表している。二次DASD416に更新内容を れる、滋糜データ・シャドーイング・プロセスのバック **郡き込む際に入出力エラーが発生するか、または二次側** 

プロセスに入っていた整合性グループ全体を最初から再 16の入出力が行われたか、どの入出力が行われなかっ たか、どの入出力が処理中かなどを追跡せずに、遠隔シ 窗用することができる。このため、どの二次DASD4 ヤドーイングを行うことができる。 [0061] 二次入出力書込み

ボリュームでの発生順に維持されている限り、その入出 ステップ1160の重要な構成要素は、二次側431が によってレコードが二次DASD416に効率よく告き 込まれることである。必要な効率は、主に、様々な二次 DASD416への複数の入出力操作を同時に実行する 21からかなり遅れてしまう恐れがある。単一チャネル ・コマンド・ワード (CCW) 連鎖を介して単一の二次 ば、二次側431ではさらに高い効率が得られる。それ ぞれの単一CCW連鎖内では、そこで行われる各二次D ASD416のデータ・トラックへの入出力操作が一次 -次側421か6後れをとらないように、PDM414 ことによって達成される。二次DASD416を一度に 1つずつ連続して書き込むと、二次側431は一次側4 技置宛ての整合性グループごとにレコードを書き込め 力操作をさらに最適化することができる。

【0062】 特定の整合性グループ用の二次入出力操作 み入出力操作のパターンに基づいて行われ、一部は二次 DASD416の物理特性に基づいて行われる。 坂適化 は、二次DASD416がカウント/キー/データ(C を単一CCW連鎖内で最適化する場合、一部は一次哲込 KD) か、拡張カウント/キー/データ (ECKD)

か、固定プロック方式 (FBA) かなどに応じて、多少 に100-次DASD406に対して行われる複数のW RITE入出力 (m) は、1つの二次DASD416の ボリュームに対する単一のSTART入出力操作に削減 することができる。このように二次記憶制御装置415 と、二次DASD416は後れをとらずに済み、それに 変化する可能性がある。その結果、所与の時間間隔の間 こ対するSTART入出力の回数をm:1に最適化する たり、一次側421のフコード更新をもっと精密にシャ ベーイングすることができる。

こよる二次入出力の最適化の重要点は、一貫性のあるコ 対して同時に行う複数の入出力操作のいずれかで発生す 書込みを伴わずに記録される恐れがある(たとえば、実 祭にはデータベース用の実際の更新書込みが失敗に終わ と示すログ項目は、二次DASD416のコピーの順序 【0063】正常な遠隔データ・シャドーイングとそれ ピーを彼旧に使用できるように、二次DASD416に る回復不能エラーを最小限にすることである。所与の二 **吹書込みで失敗すると、その後の従属書込みが条件付け** っているのにデータベース・レコードが更新されたこと 8合性に違反する)

[0064] その更新失敗が復旧されるまで、失敗した 二次DASD416のコピーはアプリケーションの復旧

ಜ

431全体が停止し、最初期設定された場合は、巷込み

(14)

特開平7-239799

こCWは、トラックの残りを消去してレコード1を更新 するため、レコード5は削除される。この連鎖を再実行 しなければならない場合、レコード5の先頭に配置する LOCATERECORD CCWが位置次めポイント 吹側421ですでに書込み操作が正常に行われているの で、データの整合性と保全性を維持するには、二次DA SD416上の整合性グループ全体をいつでも適用でき を持たなくなる(レコード5が存在しなくなる)ため、 この連鎖は先頭から完全に回復することができない。

する。第一のカテゴリは、同じ二次DASDボリューム レコードの入出力命令を含む (すなわち、同じDASD 図11のステップ1160によって表され、図12に定 義されるFCGR規則を使用するプロセスの詳細を示す ものである。ステップ1410でSDM414は、現行 整合性グループの各種レコードを2つのカテゴリに分割 向けの入出力命令を含み、第二のカテゴリは、第一のカ テゴリに含まれるレコードのうち、同じCCHH向けの 【0068】図16のステップ1410~1470は、

ることが必要である。

【0069】現行整合性グループのレコードのカテゴリ 分類後、ステップ1420では、トラック上のデータ配 個を職別し、トラック/レコードのアドレス指定を行う ために、アプリケーションのWRITE入出力およびS DM414のWRITE入出力を二次DASD416の 方式、たとえば、ECKD方式のFCGR規則(図12 を参照)に適合させる。SDM414は、ステップ14 TE入出力操作を単一入出力CCW連鎖にグループ分け する。ステップ1440は、実際の二次DASD416 の書込み用の検索引数と特定のレコード・データ(CK Dフィールド) に応じて、それぞれの二次DASD41 6のヘッド・ディスク・アセンブリ (HDA) を移動さ 30で、同じボリュームに対する「次DASDのWR」 トラックに更新されるレコード)。 せることを含む。 ຊ

【0070】ステップ1450では、後続の書込み操作 FERS1と2を比較する。FCGR規則に従うと、エ によって、前の書込み操作またはDASD検索引数(こ こで消去されるレコードでの分割など)を無効にするか どうかを判別するために図12のFCGR規則を使用し て、第二のカテゴリ(通常、レコードを受け取るトラッ を構成するレコードについてREAD SET BUF ラーが発生した場合に、一次側421からレコード更新 を再度受け取らなくても、SDM414は整合性グルー プ全体を再書込みできるようになる。SDM414が現 ステップ1460では、状態テーブル (図1) とマスタ クごとに1つずつ、複数の第二のカテゴリが存在する) 行整合性グループを二次DASD416に適用した後、 · ジャーナル (図8) を更新する。

(現行整合性グループになるもの) を獲得して、処理を 【0071】ステップ1470は、次の整合性グループ

23

よって更新されるが、FORMAT WRITE入出力

5. Va-K5HUPDATE WRITE CCWE

に使用できなくなる。失敗した更新は、S DM4 1 4 に よって b DM4 0 4 から現行コピーを要求することで修 それ以前の他のすべての更新がPDM404によっ て処理されるまで使用できなくなる。通常、失敗した更 新の復旧に要する時間は、十分な災害復旧保護のために 正できるはずである。その間に二次データ・コピーは不 整合になり、そのため、 P D M 4 0 4 が現行更新で応答 【0065】効果的な二次側431の入出力最適化は、 受け入れられないほど長い非復旧ウィンドウを示す。

所与の整合性グループについて書き込まれるデータ・レ コード・セットを検査し、ECKD方式などの二次DA SD416の特定の方式の規則に基づいて連鎖を構築す ることで実現される。ここに開示する最適化技法は、入 出力エラーが発生した場合に整合性グループを適用する 際に、CCW連鎖を再実行できるように、または、二次 タを紛失せずに整合性グループ全体を再適用できるよう [0066] 図12は、ECKD方式用のすべてのWR **初期プログラム・ロード(IPL)復旧の場合に、デー** に、入川力エラーからの復旧を単純化するものである。

I TE入出力の組合せに対応するCCW連鎖を構築する SDトラックに対して行われるWRITE入出力操作の ための完全整合性グループ復旧(FCGR)の規則を要 約して示すもので、ここではCCHHRレコード形式が 可能な組合せをそれぞれ検査することによって作成され る。図12のFCGR規則 (図14および図15に記載 する)は、整合性グループを適用する際のエラーについ 号)。図12は、1つの整合性グループの範囲内でDA て完全復旧を行うためにデータ配置 (一次DASD41 使用される(シリンダ番号、ヘッド番号、レコード番

6の入出力各込みCCW連鎖)を管理する場合に従うも のである。図12に示すFCGR規則は、新しいWRI R規則は、同一DASDトラック分析に対するREAD レコード・セットを、一次DASD406のWR1TE 入出力タイプ、検索引数、およびカウント・フィールド TE入出力操作が追加されるたびに適切に拡張されるこ とになる。これらの規則は、二次側431のハードウェ アまたはソフトウェアで実現することができる。F C G とキー・フィールドの検査に還元するので好都合であ 【0067】図12に示すように整合性グループの書込 と、前に書き込まれたデータ・レコードの再審込みがで きなくなる可能性がある。たとえば、以下の内容の連鎖 み操作を検査せずにDASDトラックが書き込まれる があると想定する。

この場合、レコード1とレコード5は同じDASDトラ ックに存在し、レコード1がレコード5の前に置かれ U⊐-F5~ØWRITE UPDATE V=-F1~0FORMAT WRITE

二次DASD416に適用することができる。一次側4 21が補捉したデータのうち、SDM414によって完 リアルタイムで統行される。一次側421から二次側4 3 1 への通信が終了した場合、遠隔コピー・プロセスは た場合(災害が発生した場合)、規則的な運転停止が行 われた場合、または二次側431で特定の引継ぎ処置が ステップ1410に戻すので、遠隔コピー・プロセスは 停止する。この通信は、ボリューム対がPDM404に よってプロセスから削除された場合、一次側が破壊され 行われた場合に終了することがある。 二次側431でジ ャーナル処理された整合性グループは、引継ぎ操作中に 全に受け取られていないデータだけが紛失する。

စ္က に複数の光ファイバ・リンクを介して複数の一次記憶制 った場合に二次側がリアルタイムのデータ可用性を提供 【0072】 要約すると、これまで同期および非同期遊 隔ゲータニ重化システムについて説明してきた。 同期遠 ム災害復旧機能を提供するが、この遠隔データ二重化シ ステムは一次側と二次側を含む。二次側は、データ・シ ャドーイングのために一次側から書込み入川力更新を同 期的に受け取るが、災害によって一次側が使用不能にな **たきるように、二次側は一次側から数下キロメートル**離 れた位置に配置される。一次側の一次プロセッサは複数 のアプリケーションをサポートし、この複数のアプリケ ・ションによって告込み入出力更新が行われる。複数の **書込み入出力更新が複数の一次記憶装置に書き込まれる** ようにする。「孜側の一次プロセッサは、データ・リン クによって一次プロセッサに連結され、一次側の複数の 一次記憶制御装置は、一次プロセッサに連結され、さら 御装徴に連結されているため、複数の二次記憶制御装置 は、告込み入出力更新をシャドーイングするために記憶 複数の一次記憶制御装置から受け取った順に複数の二次 隔データ二重化システムは、記憶域ベースのリアルタイ - 次記憶制御装置がこの魯込み入出力更新を受け取り、 城ベースの二重対を提供する。複数の二次記憶装置は、 記憶制御装置から書込み入出力更新を受け取る。

ぎ、エラー回復を行いながら、障害が発生した二重対の 【0073】一次プロセッサ内の入出力エラー回復プロ **グラムは、複数の二次記憶制御装置と通信し、二重対の** 障害が発生した場合は、この入出力エラー回復プログラ 特徴に関して一次プロセッサと複数の二次記憶制御装置 回復プログラムは、エラー回復が失敗に終わった場合に **一次および二次プロセッサに状況を出力しながら、さら** に二重対障害発生メッセージを出力する。 入出力エラー ムが前記複数のアプリケーションから制御権を引き継 こ入出力書込み更新が行われるのを防止する。

次プロセッサとの間、および一次記憶制御装置と二次記 【0074】特に本発明の実施例に関連して本発明を図 示し説明してきたが、当業者は、本発明の精神および範 囲を逸脱せずに形態および詳細の様々な変更が可能であ **ることに留意されたい。たとえば、一次プロセッサと二** 

た、同期遠隔二重化モードの一次側と二次側を分離する 距離は、本発明の数示から逸脱せずに今後の本発明の実 **極例における改良型の通信リンクによってさらに延長す 憶制御装置との間の通信リンクは変更可能である。ま** 

[0075]まとめとして、本発明の構成に関して以下

の事項を開示する。

[0076] (1) 一次側が一次記憶サブシステムに連 **粘された一次プロセッサを含み、前記一次プロセッサが** 一次記憶サブシステムへの入出力書込み更新の原因とな る少なくとも1つのアプリケーションをそこで実行し、

前記一次プロセッサがそこで実行される入出力エラー回 **復プログラム (入出力ERP) をさらに有し、二次側が** 二次記憶サブシステムに連結された二次プロセッサを含 イングし、前記二次記憶サブシステムが入出力ERPと 通信する、遠隔データ・シャドーイングを提供する力法 こおいて、前記力法が、(a) 前記一次記憶サブシステ ムから前記二次記憶サブシステムに入出力書込み更新を 送るステップと、(b)前記一次および二次記憶サブシ ステム間に二重対が確立されているかどうかを判別する ステップと、(c) 一次記憶サブシステムから一次プロ み、前記二次側が入出力書込み更新を同期的にシャドー セッサに障害が発生した二重対を報告するステップと、

の制御権を移転するステップと、(e) 二重対の障害の (d) 二重対の障害が発生した場合に、少なくとも1つ のアプリケーションから入出力ERPに一次プロセッサ **欠記憶サブシステムに「重対障害発生メッセージを送る** 原因を判別するステップと、(f) 入出力ERPから

(2) 二面対の状況が設定されるか回復された場合に前 記一次記憶サブシステムが更新の街込みを続行すること を特徴とする、上記(1)に記載の遠隔データ・シャド **ーイングを提供する方法。** ステップとを含む方法。

(3) 障害が発生した二重対を殺告する際に前記一次記 とがステップ(c)に含まれることを特徴とする、上記 憶サブシステムが前記一次プロセッサに入出力状況とし てチャネル終了/装置終了/装置チェックを報告するこ (2) に記載の遠隔データ・シャドーイングを提供する

出力を出すことがステップ(d)に含まれることを特徴 とする、上記(3)に記載の遠隔データ・シャドーイン (4) 入出力 ERP が一次記憶サブシステムにセンス入 /を提供する方法。

(5) 二重対の障害が発生した場合に入山力ERPによ るエラー回復を実行し、二次側でデータ保全性を維持す る、上記(4)に記載の遠隔ゲータ・シャドーイングを るためのステップ (g)をさらに含むことを特徴とす 提供する方法。 (6) エラー回復が正常に完了した後で入出力ERPか ら少なくとも 1 つのアプリケーションに制御権を返すた

20

(16)

記(5)に記載の遠隔データ・シャドーイングを提供す めのステップ(h)をさらに含むことを特徴とする、

(7) エラー回復が正常に完了できなかった場合に入出 ッセージを送るためのステップ (i) をさらに含むこと を特徴とする、上記(6)に記載の遠隔データ・シャド カBRPを介して二次側に障害発生同期遠隔コピー・メ ーイングを提供する方法。

への後続の入出力否込みがさらに防止されることを特徴 (8) ステップ (i) によって、一次記憶サブシステム とする、 F記(7)に記載の遠隔データ・シャドーイン ゲを提供する方法。

(5) に記載の遠隔データ・シャドーイングを提供する (9) 一次プロセッサと二次プロセッサとの間でエラー 回復のための通信が行われることを特徴とする、上記

(10) 一次記憶サブシステムと二次記憶サブシステム とする、上記(5)に記載の遠隔データ・シャドーイン との間でエラー回復のための通信が行われることを特徴 グを提供する方法。

サブシステムと、データ・リンクによって一次プロセッ (11) 一次側と二次側を有し、データ・バックアップ 次側を使用不能にする災害が発生した場合に、二次側が 側が一次側から十分離れた位置に配置される遠隔データ **サに連結される、二次側の二次プロセッサと、二次プロ** セッサに連結され、通信リンクを介してさらに一次記憶 ップのために記憶域ベースの二角対を提供し、受け取っ たときの特定の同期順に書込み入出力更新が書き込まれ リアルタイムのデータ可用性を提供できるように、二次 が、 書込み入出力更新を生成する少なくとも 1 つのアブ **書込み入出力更新を受け取って格納するための一次記憶** サブシステムに連結され、書込み入出力更新のバックア る、二次個の二次記憶サブシステムと、一次プロセッサ リケーションを実行する、一次側の一次プロセッサと、 二重化システムにおいて、遠隔データ二重化システム のために二次側が一次側から書込み更新を受け取り、

(12) エラー回復手段が、一次プロセッサと二次側の 両方に二重対障害発生モードを報告することを特徴とす る、上記(11)に記載の遼隔データニ重化システム。 回復手段とを含む遠隔データ二重化システム。

内にあって、二次記憶サブシステムに連結され、二重対

の障害が発生した場合にエラー回復を行うためのエラー

を静止することを特徴とする、上記(11)に記載の遠 (13) エラー回復手段が、二重対の障害の原因を判別 しながら、一次プロセッサで実行中のアプリケーション 隔データ二重化システム。

ンス入出力を出すことを特徴とする、上記(13)に記 (14) エラー回復手段が、一次記憶サブシステムにセ 散の遠隔データニ重化システム。

20 (15) エラー回復手段がエラー回復を完了し、二重モ

特開平7-239799

サで実行中のアプリケーションに制御権を返すことを特 ードが再確立された後、エラー回復手段が一次プロセッ

徴とする、上記(14)に記載の遠隔データ二重化シス

ない場合に、エラー回復手段が一次記憶サブシステムの (16) エラー回復手段がエラー回復を正常に完了でき **容込み入出力更新をさらに防止することを特徴とする、** 上記 (15) に記載の遠隔データ二重化システム。

(17) 一次プロセッサに連結され、エラー回復手段か と、二次プロセッサに連結され、エラー回復手段からエ **らエラー・メッセージを受け取るための一枚コンソール** ラー・メッセージを受け取るための二次コンソールとを さらに含むことを特徴とする、「記(11)に記載の遠 隔データ二重化システム。 9

(18) 一次および二次記憶サブシステムがそれぞれ複 数の記憶制御装置を含むことを特徴とする、上記 (1 1)に記載の遠隔データニ重化システム。

(19) 通信リンクが、複数のエンタープライズ・シス テム接続(ESCON)リンクを含むことを特徴とす 8

ションを実行する、一次側の一次プロセッサと、番込み (20) 一次側と二次側を有し、データ・シャドーイン グのために二次側が一次側から書込み入出力更新を同期 的に受け取り、一次側が使用不能になった場合に、二次 二次側が一次側から十分離れた位置に配置される、記憶 域ベースのリアルタイム災害復旧機能を提供する遠隔デ **ータニ重化システムにおいて、該遠隔データニ重化シス** テムが、書込み入出力更新を生成する複数のアプリケー る、上記(18)に記載の遠隔データ二重化システム。 側がリアルタイムのデータ可用性を提供できるように、 入出力更新を受け取るための複数の一次記憶制御装置

装置と、データ・リンクによって一次プロセッサに連結 数の一次記憶制御装置に連結され、書込み入出力更新を と、甚込み入出力更新を格納するための複数の一次記憶 シャドーイングするために記憶域ベースの二重対を提供 装置に連結され、そこに書込み入出力更済を格納するた 御装置に二重対障害発生メッセージを出力し、エラー回 される、二次側の二次プロセッサと、二次プロセッサに 連結され、さらに複数の光ファイバ・リンクを介して複 次側の複数の二次配憶制御装置と、複数の二次配憶制御 て、複数の二次記憶制御装置に連結され、二重対の障害 が発生した場合に前記複数のアプリケーションから制御 権を引き継ぎ、エラー回復を行いながら、二重対の障害 の特徴に関して一次プロセッサおよび複数の二次記憶制 --次および二次プロセッサに状況を出力するためのエラ 復が失敗に終わった場合にさらに入出力更新を防止し、 めの複数の二次記憶装置と、一次プロセッサ内にあっ し、普込み入出力更新が生成された順に提供される、 စ္က

[0077]

-回復手段とを含む遠隔データニ重化システム。

【発明の効果】本発明の実施により、一次側で行った巷

る。 【図13】図12のテーブルで使用する規則の説明の構 読取りレコード・セット・コピーを書き込む方法の流れ 特累平7-239799 [図14] 図12のテーブルで使用する規則の説明の一 【図16】 完全整合性グループ復旧機能を持つ二次側に 9 エンタープライズ・システム接続(ESCON)リ 【図15】図12のテーブルで使用する規則の説明の一 \* ループ復旧規則アプリケーションを示すテーブルであ ホスト間通信リンク 10 災害復旧システム 一次記憶制御装置 二次記憶制御装置 一次プロセッサ 二枚プロセッサ 対等通信リンク 成を示す図である。 **АНЛЕ R P** 一次DASD 二次DASD チャネル ナヤネブ [符号の説明] 部である。 部である。 図である。 1 1 12 3 4. 15 7 8 읔 【図9】整合性グループを形成するためのシーケンス例 20 シーケンスの場合のECKD方式装置用の完全整合性グ\* るみ更新を二次側にシャドーイングするための改良され 【図1】同期遠隔データ・シャドーイング機能を有する 【図4】非同期遠隔データ・シャドーイング機能を有す 【図5】図4の一次側からの読取りレコード・セットの 【図8】図4の二次側が使用するマスタ・ジャーナルで [図10] 整合性グループを形成するために情報および 【図11】整合性グループを形成する方法を示す流れ図 [図2] 図1の災害復旧システムにより同期遊隔コピー 【図7】ボリューム構成情報を識別する状態テーブルで 読取りレコード・セットを収集する方法を示す流れ図で 【図12】 DASDトラックに対する所与の入出力操作 【図6】読取りレコード・セットを構成する各種フィー [図3] 入出力エラー回復プログラム (入出力ERP) 前に付く接頭部ヘッダを示すデータ形式図である。 た設計および方法を提供することができる。 る災害復旧システムのブロック図である。 災害復旧システムのブロック図である。 を提供する方法を示す流れ図である。 操作の方法を示す流れ図である。 ルドを示すデータ形式図である。

800 42*4* · 94-11

「ソフトウェア製会性グループ」 別にグループ分けされた

教室のレコード

[図13]

**阿** 

E 1 2

取合性グループ番号 ジャーナル上の位観 抜作タイム・スタンプ

[<u>8</u>8]

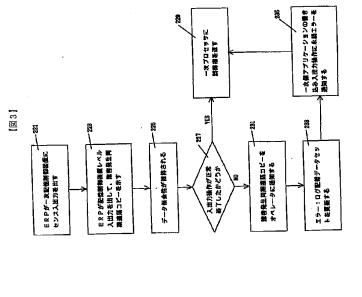
[図 図

二分的态形的数值 ニオブロセッサ 11 A C 의 入出力 ERP 一分記載起間兼信 ーオプロセッサ DA SD

梅暦平7-239799 (18)

(17)

<u>⊠</u>1



二次DASDにデータ 更新を審合込む

ĕ

に無対の状態が の不参ざがが

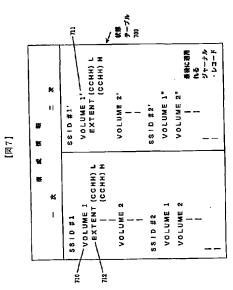
Ĭ,

・ 大記様知識を設定でチーク更新を 送信する 一次DASDにデーク製剤を整合 込み、二次DASDに搭摘する - 213

エラー回復プログラムが製御権を 引き絡ぐ

ផ្ល

に開発の政権を一次プロセッサに 数値する 一次因権政務政権が入出力状況と してOE/DE/教師チェックを 発信する

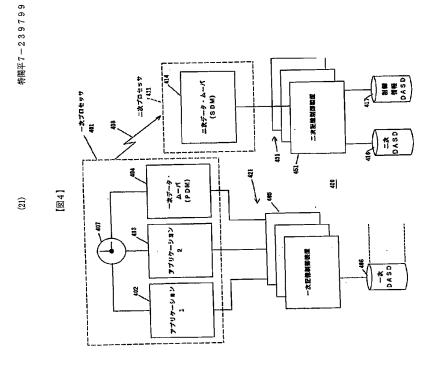


特開平7-239799

(19)

[図2]

報題



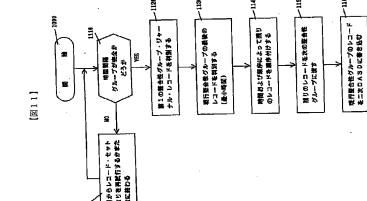
[图]

特開平7-239799

0.00	存在のイヤ・マウング	島西部屋	iek	国際母国ノ国和級国	
		糠	原序3の1	原件3の2	<b>原第303</b>
SSID1	F	10	11:59 ②	12:00 G	12:01
58:02	ī	5	12:90 @	12:02 Ø	
58   03	f	1.5	11:58 ©	0 85:11 0	12:02 @
SSIDI	72	. ZI D			
50155	_2_	8			
82128	72	8			
SSIDS	Т-	<u>8</u> -			
		<u> </u>			
整合性がル	整合性グループ条件1				
£ 11:58			19年の第7年の本の本の本の書の書の記して	の事を与い権力	1 三
Ø 11:59		8 S	SSIDを存む場合はに登録の記	の日本中に	.
9 11:59					
Ø 12:00					
G 12:08					
12:01		100	SSIDを食べの最大規略等間の最小時間	過剰を図り扱う	里
整合性グル	数合性グループ番号2				
6					
6					

フリード・カットの経行くシグ 例本ない アコード・カット会会や送板 すめ 基数し行フリード・カット編 海和器砂製や 時間間隔パンに二次銀でレコードも収集する ジャーナル発掘したレコード や二次銀に活躍する 申1次DASDボリュームと のセション表面差する 数会性グループ条形成する アプリケーションス出力操作 にタイム・スタンプを付ける NO レコード・セット 信義な影会かどうか [図10] (24) 女権アコード名群送信する よら一次銀に開発する

特開平7-239799



_		
_		
•		
-		
•		
<u>-</u>		
2		
-		

(36)

2
-
図

			<u> </u>			LYG	<b>心計載存近書</b> 代	ΥĦ				
KC≄0 化G键键型	化学を設定 化学を対象	去軟仓聯	去配金表	ー キ C 未額 応収報ようか	ーキ C 会院 ゆび替べいア	砂道を受容 のキコソ	・ 野谷 野子 野子 大 に の に る に 。 に る に 。 に る 。 に 。 に 。 に 。 に 。 に 。 に 。 に 。 に 。 に 。 。 に 。 に 。 に 。 に 。 に 。 。 。				_	
,3	W	К	O	١	N	,3	,M	蜂	¥	<b>€</b>		
. W	, B	Ж	a	r	И	<u>,</u> ^	.=	#	Ħ		₹	4
Я	8	T	ય	T	a	1		48				2 ₹
, M	M	٦	4	н	N	0	٥	<b>44</b>	<del>*</del> 4 €	<u>. 7 - </u>	- # 4	: 46
1	T		ъ	. ષ્ટ	સ .	T	7	¥	K		¥	а
M	<b>M</b>	Ð	3	W	N	8	8	¥	K		#	8
	w	м	3	м	M ·	*3	W	4	-	)=7 ¥		Ŧ
M	, a	M	3	<b>"</b>	Α	w	,=	н	ন	# 7	3	1

[図14]

B -参考1の検察が表考2の検索引数より大きい場合、参与1を捨てるが関方とも実行する

0.珍多1のフリードが参考2万路1のフリードと等しいがそれ以上でいるも後も、推布1位指行ら か両方と名案行する

D -春号2がROを見振している場合、西方とも責行するかエラーになる

E‐エラー (発生してはならない)

\*・参考1と参与2が高じフコードもある機会、エサーに対象 (両値関わ形式等3多を行むずに解出してはならない)

F・番号2の第1のレコードがR1である場合、両方とも巻き込むか、エシーになる

G -動作1の後輩引数が番号2の後非引数と移しいかぞね以上である場合、番号1を換てるかエラー

H - 春号 1 0 後後引致が参与2 0 の数数の レコードより大きい場合、参与1を指てる。参允は乗号2 D 後限引数が参与1 0 後級の レコードより大きい場合、エラーになるが両方とも寄むこむ

(車車)の装置引数が着条2の最後のフェードの参与いが木に対したある) 斧状状 (車車)の前間のフェードが車を2の最後のフェードが車にかられば上である) レジル (車車2の設度型数が振手)の自殺のフェードが降しいが、それ以下である) 独全書手 (会家でもおされ、後来なるのと上下と終しいが、それ以下である) 独全書手 (会家にもおされ、毎年を1の表達してある) 地上になる。 第4年 (金米のの実施上数が車車1の会議しのフェードとび大手に) × -かの戸敷殖化するにはな下の手腕を実行いむる

エラーになる または関方とも着き込む

(38)

特開平7-239799

[図15]

」 - 連巻2のレコード(または装痕型数)が参考1の発験のレコードより大きい場合、エサーだなる

か関方とも繋ぎ込む

スーを参えのフリード(単位は後後出版)が着ちこの故意思教より大台に始合、ゴリーにならが国だ とも参考込む

1、一番も1の独展引数が撤去2の後週引数と等しいかそれ以上である場合、同方とも総会込むかまた は参与しを捨てるかあるいはエラーになる

8 - (森布1の故景処教が落本2の後接別数が移じいかれた上である) 場合、参与1を放てる

家たは両方とも審会込む

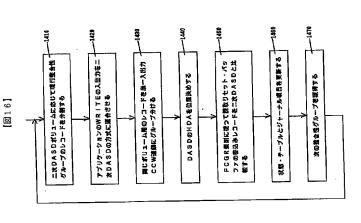
14. 巻与2の後級引盤が巻号1の反驳のレコードより大きい巻合、エラーになるが関方とも書き込む

R -春号 1を捨ててもよい

T-番号1を替てなければならない

W-週分とも奇ながけ

2、最も1人指数2岁度ワフロード供給の組合、適多1条指にながれた原於力の保たするがまれるが、最も1人指数2の方式によりの表別の表質にある。まれはフロードを関係のなってしての表別の表質にある。



フロントページの統含

(72)発明者 マイケル・アロイシウス・ポールソン (72)発明者 ウイリアム・フランク・ミッカ アメリカ合衆国85718 アリゾナ州ツーソ

ン イースト・ラエスパルダ 3921 クラウス・ウィリアム・ミッケルセン アメリカ合衆国95120 カリフォルニア州 (72) 発明者

サンノゼ シーナリー・コート・ドライブ 6567

(72)発明者 ロバート・ウェズリー・ショムラーアメリカ合衆国96037 カリフォルニア州モーガン・ヒル ピエドモント・コート17015

モーガン・ヒル ホウィッパーウィル・ド アメリカ合衆国95037 カリフォルニア州

ライプ2901